DEVELOP DIVERSE™

AI Bias in Hiring: 1439 Job Adverts Analyzed

Updated Study with New Generative AI Models

Study Outcomes

In a study of 1,439 job advertisements, GPT-4 generated ads were found to be 29.3% more biased overall compared to human-written ads. AI-generated ads scored an average of 40.9 on the inclusivity scale, while human-written ads averaged 57.9. The most significant disparities were observed in language related to neurodivergent individuals, disabled candidates, and older job seekers.

Both AI and human-written ads showed the highest inclusivity scores for male candidates, suggesting a general bias towards male-oriented language. The study also revealed that both types of ads struggled most with inclusivity for disabled and neurodivergent candidates. While GPT-4 showed some improvement compared to earlier AI models, this was largely due to a decline in the inclusivity scores of human-written ads rather than significant AI enhancements.

1439 Job Ads Analyzed

29% More Biased than Human Written Job Ads

42.5 Lowest Inclusivity Score in a Category

7 Point drop for Human Written Job Ads

Background of the Study

Category bias	Inclusivity score: Human written ads	Inclusivity score: ChatGPT written ads	GPT increase in bias – relative to human written ads
	Inclusivity score*		
Total	65	39	40%
Male	92	84	9%
Female	66	39	41%
Non-conforming	68	42	38%
Advantaged ethnicity	77	57	26%
Disadvantaged ethnicity**	65	38	42%
Young	76	54	29%
Old	67	42	37%
Abled	76	54	29%
Disabled	63	37	41%
Neurotypical	76	54	29%
Neurodivergent	63	37	41%

In February 2023, we conducted a study to understand how biased ChatGPT written job ads were. There we analyzed 7000 job ads and found that GPT created 40% more biased job ads than human-written job ads.

How did we conduct the study?

Accumulating Publicly Available Job Ads

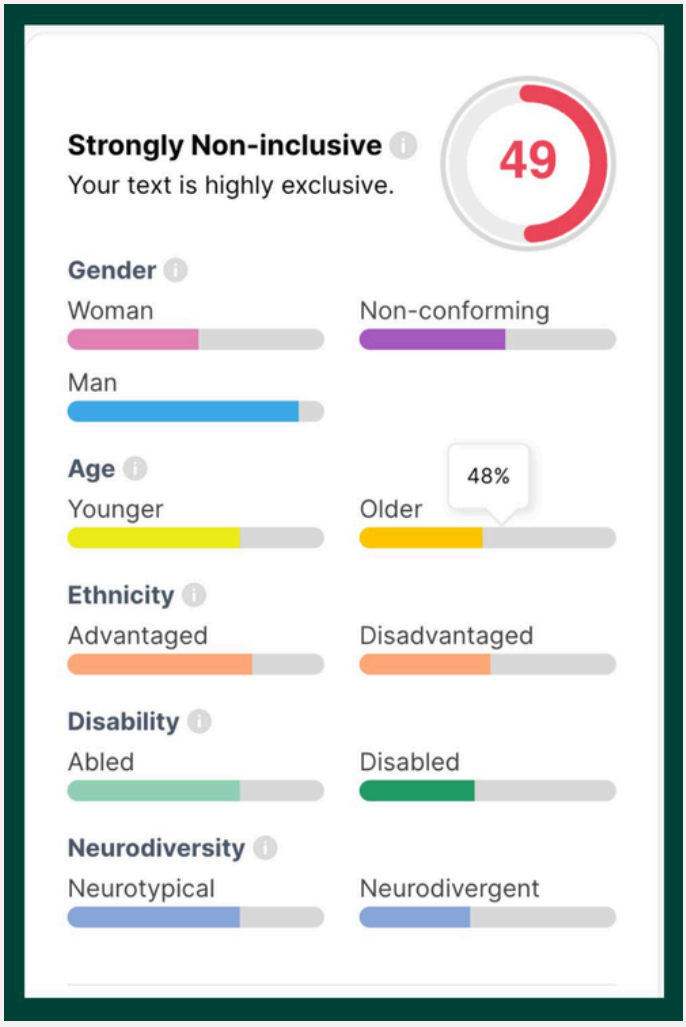
We accumulated publicly available job adverts from the internet. Next, we tokenized the job adverts to filter the job titles, companies, industries, etc. details useful for generating the prompt. At the same time, we ran job adverts through our platform to get their detailed inclusivity scores.

Next, our linguist, Linea Almgren, instructed GPT 4 to make it understand the concept of diversity, equity, and inclusion and prompted it to create inclusive job ads. After this, we made the same-titled job adverts using the prompt of Linea with GPT 4.

Finally, we ran the GPT 4 generated job ads through our platform and compared the score with the human written ones.

The Numbers

Our platform analyzes text based on different categories of bias and assigns scores based on the perceived level of bias. In other words, our software can detect what words and phrases affect specific groups. In the snapshot, the sample text is given an inclusivity score of 49 (we recommend a score above 90) and scores 48 in the 'Older' category.



How Biased are GPT 4 Generated Job Ads?

Score Type	GPT-4 Average	Real Average	% Increase in Bias
abled	68.11	76.38	10.83
advantaged	68.13	76.64	11.09
disabled	42.47	56.67	25.05
disadvantaged	46.44	60.68	23.47
female	48.06	61.26	21.55
male	88.35	93.72	5.73
neurodivergent	42.72	57.31	25.47
neurotypical	68.11	76.38	10.83
non_conforming	53.68	66.24	18.96
older	47.06	61.33	23.27
younger	68.11	76.38	10.83
overall	40.88	57.84	29.33

How to read the table?

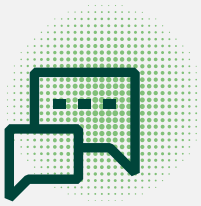
1. Score Type: This column lists various categories of potential bias in job ads. ‘Overall’ indicates the overall inclusivity score given by the Develop Diverse platform.

2. GPT-4 Average: Shows the average inclusivity score for GPT-4 generated ads in each category. Higher scores indicate better inclusivity.

3. Real Average: Displays the average inclusivity score for human-written ads. Again, higher is better.

4. % Increase in Bias: Indicates how much more biased GPT-4 ads are compared to human-written ones. A positive percentage means GPT-4 ads are more biased in that category.

Key Findings



Overall Bias 01

GPT-4 generated job ads are 29.3% more biased overall than human-written ones, with an average score of 40.9 compared to 57.8 for human-written ads. This significant difference suggests that while AI has made strides in generating coherent job descriptions, it still lags behind humans in terms of creating inclusive content.

Consistent Bias Increase: 02

GPT-4 exhibits increased bias across all categories, indicating a systematic issue. While lower inclusivity towards the 'male' category is positive, the larger gender bias gap compared to human-written ads highlights AI bias in hiring.

Most Biased Categories: 03

Both GPT-4 and human-written ads struggle most with inclusivity for disabled and neurodivergent candidates, showing the lowest scores in these categories. For GPT-4, the scores are 42.5 for disabled and 42.7 for neurodivergent candidates, compared to 56.7 and 57.3 for human-written ads. In short, from the table, we can see that physically disabled and neurodivergent people are excluded the most.

How Much Has the GPT Improved?

GPT-4 generated job ads are 29.3% more biased overall than human-written ones, with an average score of 40.9 compared to 57.8 for human-written ads. This significant difference suggests that while AI has made strides in generating coherent job descriptions, it still lags behind humans in terms of creating inclusive content.

We have 2 points to consider here -

- Human Scores Declining: Comparing the two studies, human-written ads show a noticeable decrease in inclusivity scores across most categories.
- GPT-4 Scores Improving: GPT-4's scores have generally improved or remained similar between studies. The overall score increased slightly from 39 to 40.9.

Reflecting on this, our linguist Linea explained, "We can actually see this by comparing the table from last year to the new table. The drop in inclusivity in the human-written ads from last year to this year (65 to 58) is greater than the improvement in inclusivity in the GPT4 ads from last year to this year (39 to 40)."

This indicates that the apparent improvement in GPT-4's performance is primarily attributed to the decline in human-written ad scores, rather than a dramatic increase in GPT-4's capabilities.

Which Generative AI Model is Most Inclusive?

ChatGPT 4o

Gemini 1.5

Claude 3.5 Sonnet

Based on our study we found that Claude 3.5 Sonnet is more inclusive than GPT 4o, which is more inclusive than Gemini 1.5.

Sector	Position	Country	ChatGPT 4o Score	Gemini Score	Claude 3.5 Sonnet Score
Finance	Finance Reporting Manager	Mexico	46	35	67
Retail	Store Manager	Denmark	28	19	36
Tech	Senior Software Engineer	USA	54	27	61
Offshore Energy	Deck Electrician	Offshore	60	47	49
Journalism	Senior Editor	UK	35	25	58
Travel and Leisure	Junior Product Designer	Remote	53	25	62
Healthcare	Medical Assistant	USA	60	26	62
Automotive	Maintenance Mechanic	Germany	38	38	53
Education	History Teacher	Ireland	38	31	39
Security	Regional Security Officer	Qatar	49	49	62
		Average	46.1	32.2	54.9

Analysis Summary

Claude 3.5 Sonnet achieved the highest average inclusivity score of 54.9. GPT-4o came in second with an average score of 46.1, 16% lower than Claude. Gemini 1.5 had the lowest average score of 32.2, 41% lower than Claude and 30% lower than GPT-4o.

These scores indicate that while AI has made progress in generating inclusive content, there's still substantial room for improvement across all models. For users considering these tools, it's crucial to understand that even the best-performing model (Claude 3.5 Sonnet) is achieving only about 55% of the potential inclusivity in job ads.

Final Thoughts

Our studies have revealed significant challenges in AI-generated job ads. While GPT-4 showed increased bias compared to human-written ads, our comparison of leading AI models found Claude 3.5 Sonnet to be the most inclusive, though still far from perfect. These findings underscore that AI, while valuable, should not be used in isolation for creating job ads. Human expertise remains crucial in ensuring inclusivity. As AI evolves, we can expect improvements, but a balanced approach combining AI assistance with human judgment is currently essential.

To address these challenges and create truly inclusive job advertisements, tools such as Develop Diverse can play a crucial role. Our platform helps identify and eliminate bias in job ads, whether human-written or AI-generated. By leveraging such tools and maintaining a commitment to inclusivity, we can work towards a more equitable and diverse job market for all.

Get Started

Connect with Us Today

Our website	www.developdiverse.com
Our e-mail	contact@developdiverse.com
Address	Vesterbrogade 149, 1620 Copenhagen Denmark